*Mathematical Modelling*, Vol. 7, pp. 1365–1373, 1986
Printed in the U.S.A. 

0270-0255/86 $3.00 + .00


# REGRESSION MODELS FOR COHORT MORTALITY STUDIES*,††

ALICE S. WHITTEMORE

Department of Family, Community and Preventive Medicine
Stanford University School of Medicine
Stanford, CA 94305, USA

(*Received* 14 *January* 1985; *revised* 14 *June* 1985)

**Abstract**—Cohort studies evaluate suspect health hazards from occupational or environmental exposures by recording the facts and causes of deaths in the exposed group as they occur over an extended time period. This article reviews several methods for analyzing cohort mortality data and shows them to be special cases of a single procedure. The procedure represents death rates as the product of an age-specific baseline rate that applies in the absence of exposure, times a function of exposures. Maximum likelihood methods are used to estimate unknown regression parameters in the function of exposures. The loglikelihood kernel for the data is shown to be that of $N$ independent Poisson variates, where $N$ is the total number of person-units of mortality observation time in the study. The expected values of these variates depend on the exposures and regression parameters. The latter can be estimated using packaged software programs for Poisson regression on any microcomputer that supports ANSI Standard FORTRAN.

## INTRODUCTION

Cohort studies of disease mortality among occupationally and environmentally exposed groups play an important role in detecting and understanding hazards to health. Suppose a group of people is employed in an occupation or is living in an environment that is suspected of causing one or more fatal diseases. A cohort study attempts to verify whether or not a hazard exists by monitoring over time all deaths in the exposed group, called the cohort. For each cohort member, a record is kept of the date when his monitoring began, and of the date and cause of his death, if he dies. When possible, cohort studies also record for each cohort member the timing and intensity of his exposure to the suspect hazards, as well as personal factors such as smoking habits and race that may influence mortality risk.

An example is provided by the study of lung cancer mortality among a cohort of 3,362 white U.S. uranium miners conducted by the U.S. Public Health Service (PHS) and the National Institute of Occupational Safety and Health (NIOSH)[1, 2]. In 1950 the PHS began collecting data on exposures to alpha radiation and tobacco and on cause of death. Mortality observation began for a miner at the date of his first medical examination by the PHS and ended at death or on December 31, 1977, the most recent date of virtually complete observation. The PHS estimated individual radiation exposure rates by measuring radon levels in mines during the period 1951–1968, and by combining these levels with individual work histories. Cigarette smoking histories were self-reported at the time

* Supported by NIH Grants ES-00070 and CA-23214 and by grants to the Society for Industrial and Applied Mathematics Institute for Mathematics and Society, from the Sloan Foundation, the Environmental Protection Agency, and the National Science Foundation.
†† Computer listing available.

of medical examination. To date NIOSH has recorded a total of 194 deaths from lung cancer among miners in the cohort.

Data analysis may elucidate the following issues: (a) What are the lung cancer risks of lifelong exposure to radiation from naturally occurring radon at levels currently found in homes and office buildings? (b) What are the lung cancer risks of underground uranium mining at the current U.S. Federal Standard? (c) Do cigarette smokers have radiation-induced lung cancer risks different from those of nonsmokers? (d) How do radiation and cigarette smoke induce lung cancer?

The uranium miner study illustrates several features that complicate the analysis and interpretation of cohort data. These include long and variable time periods between exposure onset and disease occurrence; exposures which vary with time, permit measurement only with error, decrease or end due to disease occurrence and correlate with other disease predictors such as cigarette smoking, ethnicity and year of birth; mortality observation periods which vary from subject to subject and which overlap exposure periods; censoring due to loss to followup or death from causes other than the one of interest; and a possible "healthy worker effect" characterized by lower disease risk among employed groups than among general population comparison groups.

These complications have motivated several methods for analyzing such data. Here we discuss methods based on a model proposed by Cox[3]. It assumes that a subject's death rate is the product of two factors. The first factor, called the baseline death rate, represents his age-specific death rate in the absence of exposure. The second factor, called the relative risk function, depends upon his exposures. Interest centers on estimating the parameters in this function. The baseline death rate may be known or it may be estimated from the data, or left unspecified.

This paper shows that many of these methods are special cases of a general maximum likelihood procedure. The procedure represents the loglikelihood kernel (i.e. those summands in the log of the probability for the data that depend on unknown quantities) as that of $N$ independent Poisson variates, where $N$ is the total number of person-units of observation time contributed by all subjects in the study. One then partitions these person-units into subsets over which exposures are assumed constant. When the baseline death rate is unknown it is approximated by a linear combination of known functions with unknown coefficients. The resulting likelihood equations for the parameters in the relative risk function are the same as those obtained using a known baseline rate, with the known rate replaced by its approximation. The equations are shown to be those of a "likelihood" function that specializes to one used extensively in analyzing failure data[3,4]. Packaged software programs can be used to estimate the parameters in the relative risk function (and if the baseline rate is unknown, the coefficients in its approximation).

## THE MODEL

Because death rates due to specific diseases vary appreciably with age in ways that are well known, we represent such rates as functions of age and take a subject's "failure time" as his age at death from the disease of interest. We assume that the $i^{\text{th}}$ subject's instantaneous probability of failing in the study of age $t$, conditional on his survival to age $t$ and on his exposure history, is

$$\gamma(t)\ \exp[\beta x_i(t)]. \tag{1}$$

In (1) the baseline death rate $\gamma(\cdot)$ is assumed to be a smooth function of age whose form may not be known. We later discuss possible dependence of $\gamma$ on fixed and age-dependent personal factors such as year of birth and smoking status. The parameter $\beta$ is

an unknown $m$-dimensional vector, $x_i(t)$ is a known $m$-dimensional function whose components summarize the $i^{\text{th}}$ subject's exposure history up to age $t$, and $\beta x$ denotes the inner product. The exponential function can be replaced by a smooth nonnegative function $\rho(\beta, x)$ satisfying $\rho(0, x) = \rho(\beta, 0) = 1$.

Each subject's data include the ages $s_i$ and $t_i$ when he starts and ends mortality observation, his exposure function $x_i$ defined from birth until age $t_i$, and an indicator $\Delta_i$ taking the value 1 if he died from the disease of interest and 0 otherwise (censoring). We assume that his time to failure and time to censoring are statistically independent random variables. Then the loglikelihood kernel of the data for the $n$ subjects in the cohort can be shown to be

$$L = \sum_{i=1}^{n} [\Delta_i \log \{\gamma(t_i) \exp[\beta x_i(t_i)]\} - \int_{s_i}^{t_i} \gamma(t) \exp[[\beta x_i(t)]dt] \tag{2}$$

([5, Chap. 5]).

We shall estimate $\beta$ by maximizing (2). To do so, it is convenient to rewrite (2) by changing the units of observation from subjects to person-units of time in the study, and from continuous to discrete time. We take the time units of equal length sufficiently small relative to the human lifespan so that $\gamma$ and the $x_i$ can be assumed constant within units and so that with some roundoff convention subjects contribute integral numbers of units (e.g. months). Each person-month $u$ of study time corresponds to a pair $(i_u, t_u)$ representing the index and age of the subject who contributed it. Let $\delta(u) = 1$ if the $i_u^{\text{th}}$ subject fails at age $t_u$ and $\delta(u) = 0$ otherwise. Use the notation $\lambda(u)$ for $\gamma(t_u)$ and $z(u)$ for $x_{i_u}(t_u)$ in (2) to obtain

$$L = \sum_{u \in U} \delta(u) \log\{\lambda(u) \exp[\beta z(u)]\} - \sum_{u \in U} \lambda(u) \exp[\beta z(u)]. \tag{3}$$

Here $U$ is the set of all $N$ person-months contributed by subjects in the study. Despite the randomness of $U$, (3) is the loglikelihood kernel for $N$ independent Poisson variates $\delta(u)$ with means $\lambda(u) \exp[\beta z(u)]$.

Differentiating (3) with respect to $\beta$ gives the $\beta$-likelihood equation

$$\sum_{U} \{\delta(u) - \lambda(u) \exp[\beta z(u)]\} z(u) = 0. \tag{4}$$

Solving (4) for $\beta$ requires assumptions about the functions $\lambda$ and $z$. Most cohort studies of mortality from an infrequently occurring disease involve several thousands or millions of person-months. Although the first term in brackets in (4) in zero for all but the few person-months corresponding to failure, summing the second term is intractable when $\lambda$ is unspecified or when $z$ assumes many values. To solve this problem, we introduce a piecewise constant approximation $\bar{z}$ for $z$ obtained by partitioning $U$ into $K$ disjoint subsets $U_k$, $k = 1, \ldots, K$, and taking

$$\bar{z}(u) \equiv z_k \quad \text{when } u \in U_k, \quad k = 1, \ldots, K. \tag{5}$$

Substituting (5) into (3) yields the approximation

$$\mathbf{L} = \sum_{U} \delta(u) \log \lambda(u) + \beta \sum_{k=1}^{K} d_k z_k^* - \sum_{k=1}^{K} \Lambda_k \exp(\beta z_k^*), \tag{6a}$$

where

$$d_k = \sum_{U_k} \delta(u) \tag{6b}$$

and

$$\Lambda_k = \sum_{U_k} \lambda(u). \tag{6c}$$

The β-likelihood equation corresponding to (6) is

$$\sum_{k=1}^{K} [d_k - \Lambda_k \exp(\beta z_k^*)] z_k^* = 0. \tag{7}$$

When $\lambda$, and thus the $\Lambda_k$, are known, solutions $\hat{\beta}$ can be obtained using iteratively reweighted least squares (IRLS) procedures [6] available in software packages (e.g. GLIM[7]). GLIM provides maximum likelihood estimates and hypothesis tests for parameters in a generalized linear model. Such a model specifies: (i) independent random variables $d_1, \ldots, d_K$ distributed according to an exponential family distribution (e.g. the normal or Poisson distribution); (ii) explanatory vectors $z_k^*$ available for each observation and describing the linear part of the model through $\eta_k = \beta z_k^*$; (iii) known "link" functions $C_k(\cdot)$ relating $\eta_k$ to the mean $\mu_k$ of $d_k$ via $\eta_k = C_k(\mu_k)$. Since (6a) is the logliklihood kernel for $K$ independent Poisson counts $d_k$ with means $\mu_k = \Lambda_k \exp(\beta z_k^*)$, one can maximize it with respect to β using the link function $C_k(\cdot) = \log(\cdot)$, with the terms $\log \Lambda_k$ specified as "offsets." The analysis can be performed on any microcomputer with software that supports ANSI Standard FORTRAN using a special program [8, 9].

GLIM can also be used to estimate β when the exponential function (7) is replaced by the more general $\rho(\beta, z_k^*)$. The Poisson means $\mu_k$ are specified by the loglinear model

$$\log \mu_k = \log \Lambda_k + \log \rho(\beta, z_k^*),$$

whose "linear predictors" $\log \rho(\beta, z_k^*)$ are nonlinear functions of the unknown parameter β. Frome[8–9] has suggested replacing $\log \rho(\beta, z_k^*)$ by its Taylor series expansion

$$\log \rho(\beta^0, z_k^*) + \delta D_\beta[\log \rho(\beta^0, z_k^*)]$$

about an initial estimate $\beta^0$. Here $\delta = \beta - \beta^0$ and $D_\beta(\cdot)$ denotes the vector of first partials with respect to the components of β. GLIM is used to estimate δ by treating $D_\beta[\log \rho(\beta^0, z_k^*)]$ as an explanatory vector and adding the known term $\log \rho(\beta^0, z_k^*)$ to the offset. One then sets $\beta^1 = \beta^0 + \delta$ and repeats the procedure until the estimates converge (see [10] for details).

When $\lambda$ is unknown the following procedure yields likelihood equations for β that are formally identical to (7) with $\lambda$ replaced by a spline function estimator. To describe the procedure let the ages

$$0 = \tau_0 < \tau_1 < \cdots < \tau_{J-1} < \tau_J = T \tag{8}$$

divide the human lifespan $(0, T)$ into subintervals $(\tau_{j-1}, \tau_j]$, $j = 1, \ldots, J$, and define

$$\tilde{\gamma}(t) = \sum_{j=1}^{J} \gamma_j \phi_j(t), \quad t \in (0, T]. \tag{9}$$

Here $\gamma_1, \ldots, \gamma_J$ represent unknown parameters and $\phi_1, \ldots, \phi_J$ are known simple functions, such as piecewise polynomials of low degree. By setting $\tilde{\lambda}(u) = \tilde{\gamma}(t_u)$ and $\varphi_j(u) = \phi_j(t_u)$ we can rewrite (9) as

$$\tilde{\lambda}(u) = \sum_{j=1}^{J} \gamma_j \varphi_j(u). \tag{10}$$

Let

$$W_j = \{u \mid t_u \in (\tau_{j-1}, \tau_j)\} \tag{11}$$

denote the set of person-months contributed by subjects while in the $j^{\text{th}}$ age interval, $j = 1, \ldots, J$. Replace $\lambda(u)$ in the first term of (6a) by $\sum_j \gamma_j 1(u \in W_j)$, where $1(\cdot)$ is the indicator function. Also replace $\lambda(u)$ in (6c) by $\tilde{\lambda}(u)$ of (10). Then (6) becomes

$$\mathbf{L} = \sum_{jk} \{d_{jk} \log[\gamma_j \exp(\beta z_k^*)] - \gamma_j \exp(\beta z_k^*)\Phi_{jk}\}. \tag{12a}$$

Here

$$d_{jk} = \sum_{W_j \cap U_k} \delta(u) \tag{12b}$$

represents the failure count occurring among person-months in $W_j \cap U_k$, and

$$\Phi_{jk} = \sum_{U_k} \varphi_j(u) \tag{12c}$$

is a measure of "effective time on test" in $W_j \cap U_k$.

Expression (12a) is the loglikelihood kernel for $JK$ independent Poisson counts $d_{jk}$ with means $\mu_{jk} = \Phi_{jk} \exp(\theta Z_{jk})$. The unknown parameter is $\theta = (\ln \gamma_1, \ldots, \ln \gamma_J, \beta)$, and the explanatory variables are $Z_{jk} = (e_j, z_k^*)$, where $e_j$ denotes the $j^{\text{th}}$ unit vector in $J$ dimensional Euclidean space. GLiM can be used to maximize (12a) with respect to $\theta$ by specifying the log as link function and the terms $\log \Phi_{jk}$ as offsets. The techniques of Frome again apply when the exponential factor in (12a) is replaced by a more general term $\rho(\beta, z_k^*)$.

It is useful to compare the resulting estimates for $\beta$ with those obtained using a known functional form for the baseline death rate. The $\beta$-likelihood equation corresponding to (12) is

$$\sum_k [d_{\cdot k} - (\sum_j \gamma_j \Phi_{jk}) \exp(\beta z_k^*)] z_k^* = 0, \tag{13}$$

where the dot notation indicates summation over the replaced index. The $\gamma_j$-likelihood equations are

$$\gamma_j = d_{j\cdot} / \sum_k \exp(\beta z_k^*)\Phi_{jk}; \quad j = 1, \ldots, J. \tag{14}$$

At a maximum point $(\hat{\gamma}_1, \ldots, \hat{\gamma}_J, \hat{\beta})$, the $\hat{\gamma}_j$ must satisfy (14). Therefore we substitute (14) into (13) to obtain the $\beta$-equation

$$\sum_k [d_{\cdot k} - \hat{\Lambda}_k \exp(\beta z_k^*)] z_k^* = 0, \tag{15a}$$

where

$$\hat{\Lambda}_k = \hat{\Lambda}_k(\beta) = \sum_j [d_j . \Phi_{jk} / \sum_l \exp(\beta z_l^*) \Phi_{jl}]. \tag{15b}$$

Equation (15a) is identical to (7) with the $\Lambda_k$ of (6c) replaced by (15b). Solutions $\hat{\beta}$ to (15a) thus depend on assumptions about the form of $\lambda$ only through the choice of approximation $\tilde{\gamma}$. At one extreme, when $J = 1$ in (8)–(10), these equations specify $\lambda$ up to an unknown proportionality constant $\gamma_1$. At the other extreme, when $J$ is large and the functions $\phi_j$ are piecewise polynomials of low degree, (8)–(10) give "nonparametric" representations of $\lambda$.

The asymptotic variances of solutions $\hat{\beta}$ to (15a) are larger than those of solutions to (7). Breslow *et al.*[11] have investigated the asymptotic efficiency of estimates $\hat{\beta}$ for certain approximations $\tilde{\gamma}$ relative to that of estimates obtained using known death rates $\lambda$.

By interchanging the order of summation one can rewrite the $\beta$-equation (15) as

$$\sum_j d_j . [\bar{z}_j - D_\beta \log(\sum_k \exp(\beta z_k^*) \Phi_{jk})] = 0. \tag{16a}$$

Here

$$\bar{z}_j = d_j^{-1} \sum_k d_{jk} z_k^*, \tag{16b}$$

is the mean exposure vector averaged over failures occurring in the $j^{th}$ age group. Note that (16a) can be obtained by differentiating with respect to $\beta$ the log of the function

$$\prod_j [\exp(\beta \bar{z}_j) / \sum_k \exp(\beta z_k^*) \Phi_{jk}]^{d_j}. \tag{17}$$

As shown below, this function specializes to the partial likelihood function introduced by Cox[3,4].

The approximation (9) suggests estimating the unknown baseline rate $\gamma(t)$ by

$$\hat{\gamma}(t) = \sum_{j=1}^{J} \hat{\gamma}_j \phi_j(t), \tag{18a}$$

and estimating the baseline survivor function $\exp(-\int_0^t \gamma(s)\, ds)$ by

$$\exp\left(-\int_0^t \hat{\gamma}(s)\, ds\right), \tag{18b}$$

where $\hat{\gamma}(\cdot)$ is given by (18a).

## EXAMPLES

Several loglikelihood kernels proposed in the literature for analyzing censored failure data arise as special causes of expression (12a) or are closely related to it. To describe them we specify for each the age division (8), the functions $\varphi_j$ of (10), and the partition $\{U_k \mid k = 1, \ldots, K\}$ of $U$.

*Example* 1. Let the $\tau_1, \ldots, \tau_{J-1}$ of (8) represent the $J - 1$ distinct ages at which

failures occur among all subjects. For $j = 1, \ldots, J$ set

$$\varphi_j(u) = 1(t_u = \tau_j), \qquad j = 1, \ldots, J. \tag{19}$$

Let each partitioning set $U_k$ contain a single person-month, $k = 1, \ldots, N$, so that exposures are assumed constant only within each person-month. Then $d_{jk}$ of (12b) indicates whether the $k^{\text{th}}$ person-month was contributed by a subject who failed at the $j^{\text{th}}$ failure age, and $\Phi_{jk}$ of (12c) indicates whether the subject was alive at the $j^{\text{th}}$ failure age. Expression (17) becomes Peto's[12] approximation to Cox's partial likelihood function[3], which is exact in the absence of tied failure ages. The cumulative hazard estimator obtained by integrating (18a) was derived by Breslow[13] and has been discussed by Holford[14].

Whittemore and Keller[15] discuss other estimators for $\beta$ based on this age division with alternatives to (19) for the $\varphi_j$, and on the sets

$$U_k = \{u \mid i_u = k\} \equiv V_k \tag{20}$$

consisting of all person-months contributed by the $k^{\text{th}}$ subject, $k = 1, \ldots, K = n$. They show that the corresponding baseline survival function estimators (18b) have the same asymptotic properties as those proposed by Cox[3] and Breslow[13], but have greater efficiency when the number of failures is small.

*Example* 2. Let $\tau_1, \ldots, \tau_{J-1}$ represent fixed ages, and set

$$\varphi_j(u) = 1(t_u \in (\tau_{j-1}, \tau_j]), \qquad j = 1, \ldots, J. \tag{21}$$

To partition $U$, let $K = l_1 \ldots l_m$ represent the number of cells in the $m$-dimensional contingency table formed by grouping each of the $m$ components of $z$ into $l_\nu$ discrete categories, $\nu = 1, \ldots, m$. Define $U_k$ as the set of all person-months $u$ such that $z(u)$ belongs to the $k^{\text{th}}$ cell of the table, $k = 1, \ldots, K$. With this choice of approximating baseline function and partition, the $\Phi_{jk}$ of (12c) equal the number of person-months in $W_j \cap U_k$, where $W_j$ is given by (11). This special case has been applied extensively to censored failure data when $\rho$ is the exponential function (e.g. Breslow *et al.*[11]; Holford[16]; Laird and Olivier[17]) and for more general relative risk functions $\rho$ (e.g. Frome[8, 18]). Unlike the random age-division procedures of Example 1 that yield asymptotically smooth estimates of the baseline survivor function as the number $J - 1$ of failures increases, this procedure yields an estimate that is asymptotically a step function with fixed cutpoints.

*Example* 3. Holford[14] assumed that both the baseline rate $\gamma$ and the exposure functions $x_i$ were constant over fixed age intervals $(\tau_{j-1}, \tau_j], j = 1, \ldots, J$, but that they vary among intervals and (for exposures) among subjects. Holford's method arises by choosing the spline function $\gamma$ as described in Example 2, and defining the $U_k$ as

$$U_k = U_{ij} \equiv V_i \cap W_j, \qquad i = 1, \ldots, n; \quad j = 1, \ldots, J. \tag{22}$$

Here the sets $V_i$ and $W_j$ are given by (20) and (11), respectively.

*Example* 4. Prentice and Gloeckler[19] grouped failure times into fixed age intervals $(\tau_{j-1}, \tau_j]$ and assumed that censoring occurs at the cutpoints $\tau_j$. Like Holford, they assuumed that the $i^{\text{th}}$ subject's exposure vector $x_i(t) = z_{ij}^*$ is constant within the $j^{\text{th}}$ age interval, producing the $K = nJ$ subsets $U_k$ of (22). The resulting loglikelihood kernel is

$$\sum_{jk} d_{jk} \log [1 - \exp\{-\gamma_j(\tau_j - \tau_{j-1}) \exp(\beta, z_k^*)\}] - \gamma_j \exp(\beta, z_k^*)\Phi_{jk}. \quad (23)$$

Here $d_{jk}$ and $\Phi_{jk}$ are given by (12b,c), (21), (22) in which failures are assumed to occur at the start of an age interval, and $\gamma_j = (\tau_j - \tau_{j-1})^{-1} \int_{\tau_{j-1}}^{\tau_j} \gamma(s)\, ds$ is the mean value of $\gamma(\cdot)$ in the $j^{\text{th}}$ age interval. The loglikelihood kernel (12) arises from (23) by substituting $y$ for the expression $1 - \exp(-y)$ within the square brackets.

## SAMPLING FROM FAILURE-FREE SETS OF PERSON-MONTHS

In Example 2 one can choose the numbers of age groups $J$ and exposure cells $K$ small enough to avoid empty sets $W_j \cap U_k$, and to restrict computer time used in summing terms $\exp(\beta z_k^*)$ in (15). In the remaining examples, however, $K$ exceeds the large number $n$ of subjects in the cohort. To reduce the number of terms $\exp(\beta z_k^*)$ in (7) and (15), include in the summations all sets $U_k$ for which $d_{\cdot k} > 0$, and a sample of the "failure-free" sets $U_k$ for which $d_{\cdot k} = 0$. For example when $\lambda$ is known, one can reduce the summation in (7) by sampling from the sets of person-months contributed by "control" subjects whose mortality observation ended for reasons other than failure. When $\lambda$ is unknown one can reduce summation of terms $\exp(\beta z_k^*)$ in (15) by sampling for each age interval $j$ from the set of indices $k$ such that $d_{jk} = 0$ and $\Phi_{jk} \neq 0$. When $\lambda(\cdot)$ and the $U_k$ are given by Example 1, the latter strategy becomes "sampling from the risk sets," discussed by Breslow and Patton[21], and used by Breslow *et al.*[11] and by Whittemore and McMillan[2] in analyzing lung cancer among uranium miners. Prentice and Breslow[20] have shown that the likelihood equations for $\beta$ resulting from sampling from the risk sets are the $\beta$-derivatives of an appropriately defined conditional likelihood.

The asymptotic efficiency loss in estimates for $\beta$ associated with sampling of this type has been discussed for special cases (Breslow and Patton[21]; Breslow *et al.*[11]; Whittemore[22]).

## STRATIFICATION OF THE BASELINE RATE

A straightforward extension of the preceding methods allows the baseline death rate $\gamma(t)$ to vary with stratified levels of fixed or age dependent personal factors such as year of birth or smoking status. This is accomplished by partitioning the set $U$ of person-months into $B$ subsets $U_{b\cdot}$, $b = 1, \ldots, B$, each containing all person-months contributed by subjects whose birth year and current smoking status fall in stratum $b$. One then replaces the loglikelihood kernel $L = L[\lambda(\cdot),\beta]$ of (3) by the sum

$$\sum_{b=1}^{B} L_b[\lambda_b(\cdot), \beta], \quad (24)$$

where $L_b$ is given by (3) with the summations taken over $U_{b\cdot}$. To deal with age-dependent exposures, partition each $U_{b\cdot}$ into sets $U_{bk}$, with $z(\cdot)$ approximated by a step function assuming the value $z_{bk}^*$ in $U_{bk}$, $k = 1, \ldots, K_b$. Then the $\beta$-likelihood equation (7) becomes

$$\sum_{b=1}^{B} \sum_{k=1}^{K_b} [d_{bk} - \Lambda_{bk} \exp(\beta z_{bk}^*)] z_{bk}^* = 0,$$

where $d_{bk} = \sum_{u b k} \delta(u)$ and $\Lambda_{bk} = \sum_{u b k} \lambda(u)$ represent numbers of observed and "expected" (in the absence of exposure) failures in the joint stratum $(b, k)$.

One can approximate unknown baseline rates $\gamma_b(t)$ by allowing the $\bar{\gamma}(\cdot)$ of (9) to vary among strata. In particular the age partition (8), the known functions $\phi_j$ and the unknown parameters $\lambda_j$ may be stratum-specific. With this formulation it is possible to check whether the exposure effect $\beta$ varies among strata, using likelihood ratio rests as described in [5].

## SUMMARY

The preceding discussion has outlined the flexibility and scope of a single procedure for dealing with data from complex cohort studies such as the uranium miner study. The procedure formalizes the traditional practice of classifying observed death counts and person-units of study time into categories defined by exposure and other variables. It also unifies many of the methods proposed in the literature. While these methods are special cases of one procedure, they can give different estimates for $\beta$ when applied to the same data set[23]. For example, when $\log \rho(\beta, x)$ is nonlinear in $\beta$, estimates for both $\beta$ and its standard error can be extremely unstable, fluctuating greatly between different methods. This instability suggests flatness of the corresponding likelihood functions. Further work is needed to understand the difficulties in assuming biologically plasuble functions $\rho$ that produce loglikelihood functions which are not convex in $\beta$. Additional topics for future work include strategies for choosing explanatory variables $x(t)$ so as to summarize age-dependent exposure histories, strategies for dealing with errors in exposure histories, and models to allow exposures to affect baseline rates in nonmultiplicative ways.

## REFERENCES

1. J. K. Wagoner, V. E. Archer, F. E. Lundin, D. A. Holaday, and J. W. Lloyd. Radiation as the cause of lung cancer among uranium miners. *N. Engl. J. Med.* **273**, 181–188 (1965).
2. A. S. Whittemore and A. McMillan, Lung cancer mortality among U.S. uranium miners. *J. Natl. Cancer Inst.* **71**, 489–499 (1983).
3. D. R. Cox, Regression models and life tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34**, 187–220 (1972).
4. D. R. Cox, Partial likelihood. *Biometrika* **62**, 269–276 (1975).
5. J. D. Kalbfleisch and R. L. Prentice, *The Statistical Analysis of Failure Time Data*. Wiley, New York (1980).
6. E. L. Frome, M. H. Kutner, and J. J. Beauchamp, Regression analysis of Poisson distributed data. *J. Amer. Statist. Assoc.* **68**, 935–940 (1973).
7. R. J. Baker and J. A. Nelder, *The GLIM System: Release 3*. Numerical Algorithms Group, Oxford (1978).
8. E. L. Frome and H. Checkoway, Epidemiologic programs for computers and calculators: use of Poisson regression models in estimating incidence rates and ratios. *Amer. J. Epid.* **121**, 309–323 (1985).
9. E. L. Frome, Poisson regression analysis. *Amer. Statist.* **35**, 262–263 (1981).
10. P. McCullough and J. A. Nelder, *Generalized Linear Models*. Chapman & Hall, New York (1983).
11. N. E. Breslow, J. H. Lubin, and P. Marek, Multiplicative models and the analysis of cohort data. *J. Amer. Statist. Assoc.* **78**, 1–12 (1983).
12. R. Peto, Contribution to discussion of paper by D. R. Cox. *J. Roy. Statist. Soc. Ser. B* **34**, 216–217 (1972).
13. N. E. Breslow, Contribution to discussion of paper by D. R. Cox. *J. Roy. Statist. Soc. Ser. B* **34**, 216–217 (1972).
14. T. R. Holford, Life tables with concomitant information. *Biometrics* **32**, 587–597 (1976).
15. A. S. Whittemore and J. B. Keller, Survival estimation with censored data. Stanford University Technical Report No. 69, Stanford, CA (1983).
16. T. R. Holford, The analysis of rates and survivorship using log-linear models. *Biometrics* **36**, 299–305 (1980).
17. N. Laird and D. Olivier, Covariance analysis of censored survival data using loglinear analysis techniques. *J. Amer. Statist. Assoc.* **76**, 231–240 (1981).
18. E. L. Frome. The analysis of rates using Poisson regression models. *Biometrics* **39**, 655–674 (1983).
19. R. L. Prentice and L. A. Gloeckler, Regression analysis of grouped survival data with application to breast cancer data. *Biometrics* **34**, 57–67 (1978).
20. R. L. Prentice and N. E. Breslow, Retrospective studies and failure time models. *Biometrika* **65**, 153–158 (1978).
21. N. E. Breslow and J. Patton, Case-control analysis of cohort studies. In *Energy and Health* (Edited by N. E. Breslow and A. S. Whittemore), pp. 226–242. SIAM J. Appl. Math., Philadelphia (1979).
22. A. S. Whittemore, Efficiency of synthetic retrospective studies. *Biometrical J.* **23**, 73–78 (1981).
23. J. Halpern and A. S. Whittemore, Analysis of occupational cohort data with application to lung cancer mortality in U.S. uranium miners, *J. Chronic Dis.* (in press)